Hierarchical representation of socio-economic complex systems according to minimal spanning trees and diagrams

Andrzej JARYNOWSKI*, Andrzej BUDA†

Keywords: Minimum Spanning Tree; complex systems; computational social systems

Abstract: We investigate hierarchical structure in various complex systems according to Minimum Spanning Tree methods. Firstly, we investigate stock markets where the graph is obtained from the matrix of correlation coefficients computed between all pairs of assets by considering the synchronous time evolution of the difference of the logarithm of daily stock price. The hierarchical tree provides information useful to investigate the number and nature of economic factors that are associated in a meaningful economic taxonomy. We extend this method on other financial markets – money exchange (FOREX) and commodity – phonographic market (where we have artists instead of stocks) and get information on which music genre is meaningful according to customers. We continue to use this method in social systems (sport, political parties and pharmacy) to investigate collective effects and detect how a single element of a system influences the other ones. The level of correlations and Minimum Spanning Trees in various complex systems is also discussed.

1. Introduction to hierarchical representation

Mathematical methods have become more and more popular and successfully applied in explanations of phenomena observed in real world social, economic and biological systems. We propose building a meaningful representation to show complex relations between agencies in various systems (Green and Bossomaier, 2000). We believe these visualization methods and their quantitative results can be exploited in research on markets and other social systems (not only in the examples we provided). In contrast to many other quantitative methods like statistical regressions and data mining procedures, hierarchical representation has a relatively transparent structure. Within the project, we also present a novel and unique approach to see change in hierarchical structure in time. In literature, there are attempts to model financial and commodity markets, but they do not succeed in social sciences and medicine. It is noteworthy

*Smoluchowski Institute, Jagiellonian University in Cracow
†Institute of Nuclear Physics in Cracow, Polish Academy of Science

that many relations discovered by our methodology have not been satisfactorily understood yet. Thus, we expect that our research project will bring better understanding of the processes and rules of the social systems evolution (Watts and Strogatz, 1998), because this type of relationships has not been sufficiently explored yet.

Recently, the knowledge of complex system tools for economy, sociology and medicine such as networks and hierarchical representation has undergone an accelerated growth. However, all models of such systems are incomplete without real data, especially register-based (Mezard et al., 1987). Complex network methods give the possibility to analyze phenomena of various origins as the social network characteristics emerge among others in the organization of the Internet, airline routing, relations between metabolites and enzymes taking part in metabolic pathways, networks of cooperation between authors or citations in scientific papers. Complex systems are natural or social systems that consist of a large number of nonlinearly interacting elements. The need to understand this phenomenon encourages cooperation between various registering institutions, which, in turn, exerts pressure on collecting data for simple analysis by many researchers who work on new models and use complex tools often taken from other disciplines. The most exciting property of these systems is the existence of emergent phenomena which cannot be simply derived or predicted solely from the knowledge of the system structure and the interactions between its individual elements. Methodology of physics proves helpful in successful issues of the properties of complex systems including: the collective effects and their coexistence with noise, long range interactions, the interplay between determinism and flexibility in evolution, scale invariance, criticality, multifractality (Oswiecimka et al., 2011) and hierarchical structure (Grabowski and Kosinski, 2004). The aim of our article is to fill the gap between social science and medical analyzes with a complex system approach (Kwapień and Drożdż, 2012) by applying meaningful taxonomy developed previously in the field of applied mathematics, complex systems and computer science. In this paper we investigate various complex systems like financial market (stock exchange: DJIA, DAX, FTSE100, WIG20, money exchange: FOREX) and commodity market (phonographic), social systems of political parties, sport (Polish Football League) and pharmacy, and focus our attention on identifying their hierarchical structure.

Tab. 1. Data structure of the presented examples.

Data type	**Time span**	**No assets**	**Signal**	**Length of series**	**Asset category**
Stock exchange	1997-2008	~ 30	Prize	~ 1000	Stock company
Money exchange	2002-2013	~ 30	Relative exchange rate	~ 1000	Currency
Phonographic market	2004-2014	30	Record sale	~ 400	Artists
Politics	2003-2014	~ 10	% of support in polls	~ 100	Political party
Football	2003-2004	~ 20	Points in game	~ 40	Team
Pharmacy	NA $\sim$ 2004	~ 10	Points in survey	~ 50	Health indicator

2. Correlation and its interpretation

Initially, we apply the correlation matrix approach to an empirical signal (Mantegna, 1999). The correlation coefficient defines the degree of similarity between the synchronous time evolution of a pair of assets, represented by their underlying value (prize, sale, points, preference, etc.). There are many measures of correlation like mutual information or Manhattan, but we choose the simplest linear one (Pearson).

$$\rho_{ij} = \frac{< Y_i Y_j > - < Y_i >< Y_j >}{\sqrt{(< Y_i^2 > - < Y_i >^2)(< Y_j^2 > - < Y_j >^2)}} \tag{1}$$

where i and j are numerical labels of assets, Y_i is the return or signal (underlying).

The returns are defined by: $Y_i = \ln[P_i(t)]$ - $\ln[P_i(t\text{-}1)]$ where $P_i(t)$ is the signal i at time t. The statistical average is a temporal average performed on all the trading days of the investigated time period. By definition, ρ_{ij} may vary from -1 to 1. The matrix of correlation coefficients is a symmetric matrix with $\rho_{ii} = 1$ and the $n(n\text{-}1)/2$ of diagonal elements characterizing the matrix completely. Every correlation from that matrix is based on two vectors containing P_i and P_j: the time series of signal i and j for every given time interval. The correlation coefficient reflects similarity between assets. The matrix of those coefficients can be used for revealing hierarchical structures in systems and finding the taxonomy that allows isolating groups of assets.

Three levels of correlations can be introduced (Buda and Jarynowski, 2010):

1. Strong (strongly correlated pair of assets) $\rho \in [1/2, 1]$;
2. Weak (weakly correlated pair of assets) $\rho \in [0, 1/2)$;
3. Negative (anti-correlated pair of assets) $\rho \in$ [-1, 0).

Tab. 2. The number of strongly, weakly and negatively correlated pairs in portfolios.

Correlated pairs	**Strongly**	**Weakly**	**Negatively**
DJIA	9	426	0
DAX	205	119	1
WIG 20	1	188	1
Football	2	60	58
Politics	0	7	8
Phonographic market	3	72	375

3. Minimal Spanning Tree and hierarchical diagrams

MST is a way to present a subgraph of a fully connected graph of correlations inside a system. It provides 'the core' of interaction between assets. Full correlation matrix

can be used in order to classify assets into clusters. The distance between assets is defined by:

$$d_{ij} = \sqrt{2(1 - \rho_{ij})} \tag{2}$$

and is associated with correlation coefficients. With this choice, d_{ij} fulfills three axioms of a Euclidean metric: $d_{ij} = 0$ if and only if $i = j$,
$d_{ij} = \mathrm{d}_{ji}$,
$d_{ij} \leq d_{ik} + d_{kj}$.

3.1. Stock exchange

We can build a Minimal Spanning Tree (MST) and hierarchical diagrams for a portfolio of assets. Let us illustrate it with the stock market index DJIA (Fig. 1C). Firstly we detect a link d for the strongest correlation (the shortest distance). For example, C-JPM is the strongest correlation (0.72) from the DJIA portfolio. Thus, we start to build the Minimal Spanning Tree from the distance $d_{C-JPM} = 0.75$. The second strongest correlation (0.68) is AXP-C and we could add this additional link ($d = 0.8$) to C. The third strongest correlation is AXP-JPM (0.65), but AXP and JPM is already connected by C. Another strongest correlation is GE-AXP (0.61) so we add this link to AXP. After joining all the 30 stocks we have the complete Minimal Spanning Tree (n-1 links) for the DJIA portfolio. It reflects sectors and subsectors of the economy in the investigated time period (according to daily stocks closing price history only). In Fig. 1 we present MST for this DJIA case as well as for a few other markets.

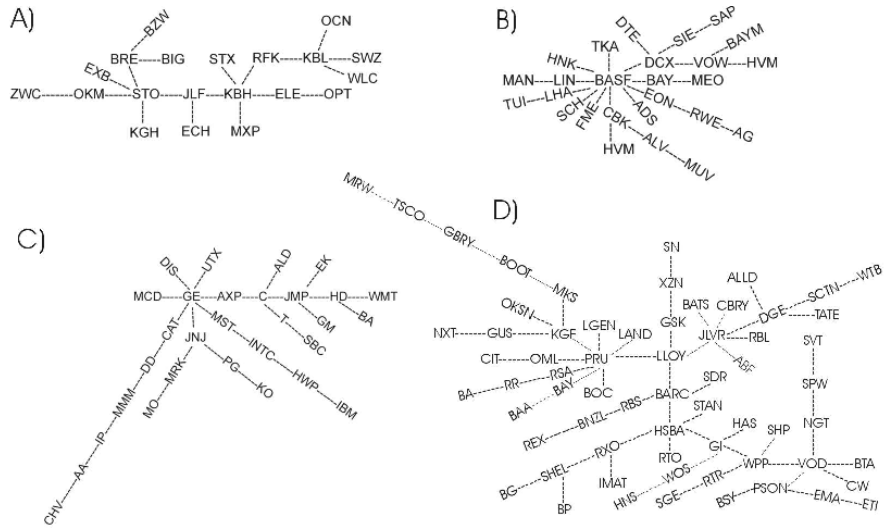

Fig. 1. Minimum Spanning Tree (MST) connecting the stocks used to compute (Buda and Jarynowski, 2010): (A) Warszawski Indeks Giełdowy (WIG 20); (B) Deutsche Aktienindex (DAX); (C) Dow Jones Industrial Average (DJIA); (D) Financial Times Stock Exchange (FTSE 100).

3.2. Money exchange

The Minimum Spanning Trees can also be obtained in the foreign exchange market (FOREX). However, the structure of these trees might depend on a reference frame, because all values $P_i(t)$ have to be expressed in their basic currency (Fig. 2). We investigated 38 currencies (including gold that had influence on all currencies strictly in the distant past) in the investigated period and detected geographical dependence between currencies.

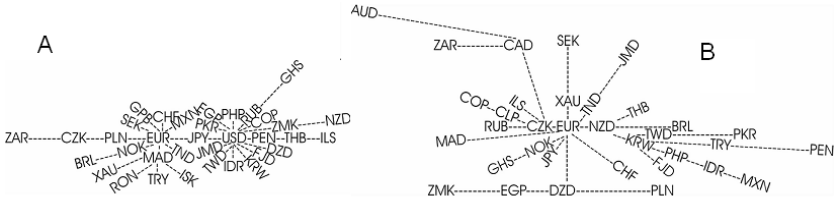

Fig. 2. The Minimum Spanning Tree obtained for 37 currencies (2002-2012). The basic currency is (A) USD; (B) AUD.

3.3. Phonographic market

Another example of a Minimal Spanning Tree (MST) comes from commodity markets, like the global phonographic market, where artists are the assets (Fig. 3). Although the hierarchical structure in financial markets reflects the Forbes classifications of stocks in the industry sectors and sub-sectors (Fig. 1), the analysis of the MST and correlations between artists does not always fit the music genres classified by the Billboard and other music magazines (Ordanini, 2006). The Minimal Spanning Tree reveals sectors that belong to rap (Kanye West, 50 Cent, Usher), rock (Green Day, Kings Of Leon, The Beatles, U2, Evanescence or Coldplay, Nickelback, Kelly Clarkson), soul and r'n'b (Adele, Britney Spears, Alicia Keys, Beyonce, Michael Jackson), but according to customers, there is no sector for pure pop music (Buda and Jarynowski, 2013).

Instead of pop, in the middle of the MST we have a celebrity sector that contains Lady Gaga, Rihanna, Bruce Springsteen, Pink, Jay-Z, The Beatles, Black Eyed Peas and Justin Timberlake. Although they represent various styles and genres, the only common thing they have is fame, high record sales, and popularity. Most of them were popular before 2003.

3.4. Politics

One of the most significant events in Poland after the Fall of Communism was the emergence of internal conflict after Polish president's plane crash on 10 April 2010. We investigate relations between political parties in Poland during the last decade. We provide analysis of preferences of voters (via polls), personal transfers between parties and public opinion associations before and after the Smolensk crash (airport in Russia where the catastrophe happened). The obtained MSTs can be viewed as representing medial 'political distances' between parties. We observe a topological

difference in MSTs before and after the critical Smolensk crash (Fig. 4). For example, polarization occurs on the PO-PiS line (these parties are no longer neighbors on the Minimal Spanning Tree, because of strong negative correlations according to opinion polls after 10 April 2010). Before Smolensk plane crash every link in the MST had a representation in transitions between the political parties, because – surprisingly – people changed political parties directly according to the MST links.

3.5. Pharmacy

The Minimal Spanning Trees can also describe social systems without given time series. For the presentation, we apply non-linear techniques like the quality of life assessment analysis made in Poznań on 86 patients using a questionnaire (WHOQOL - BREF) consisting of 26 questions (Najmrodzka, 2012). We focused on four areas: somatic (physical), psychological, social and environmental impacts, and additional

Fig. 3. The Minimum Spanning Tree obtained for the 30 most popular artists in the global phonographic market (2003-2011).

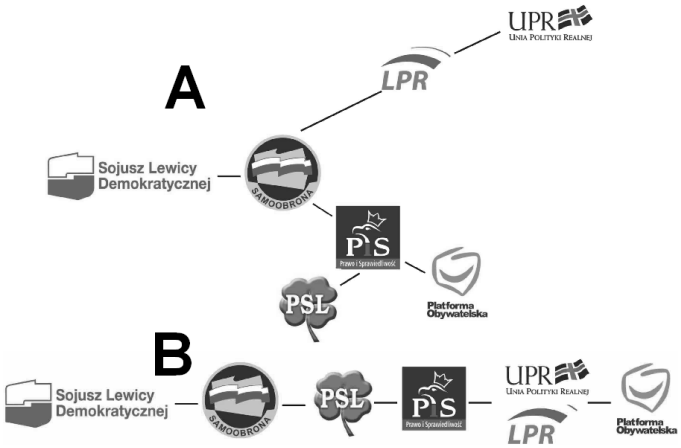

Fig. 4. The Minimum Spanning Tree obtained for Polish political parties according to opinion polls: A) 2006-2010 (before the Smolensk crash), B) 2006-2013 (after the Smolensk crash).

questions regarding the status of a self- subjective perception of health and the quality of life. Results were normalized to calculate the correlation coefficients between the investigated traits (assets). We found that the traditional medical methods of testing hypotheses receive only a few statistically significant correlations. It is a triangle: sex, weight and height, and health status coupled with quality of life. Traditional ways of presenting the characteristics that affect the quality of life are unsatisfactory. Results of multi-regression provide only a statistically significant relation between age, weight, sex and height, which are obvious. On the other hand, visualization by MST gives clarity of diagnosis: the elements that affect the quality of life directly and indirectly (Fig. 5). Thus, the MST gives new and relevant quality for epidemiological studies.

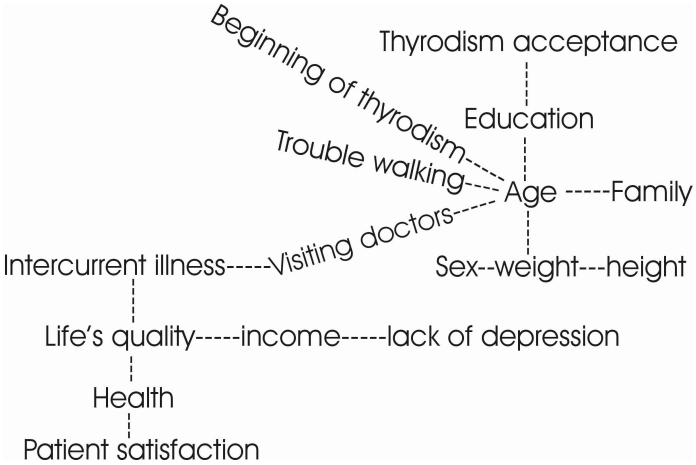

Fig. 5. The Minimum Spanning Tree obtained for the features of the people who suffer from thyroid.

3.6. Football

A subdominant ultrametric space provides a hierarchy of correlations from another perspective in a form of diagrams. The Euclidean distance d 2 is a metric. So, it is possible to draw all distances between elements starting from the shortest d and adding other ones (Fig. 6) as in Minimum Spanning Trees. We illustrate practical utility of this method by describing collective relations between Polish football clubs during the period of corruption scandals in the Polish Football League (Jarynowski, 2010). In recent years prosecution in Poland has been investigating several clubs, referees and players because of corruption procedures. We study the statistical properties of match results in the Polish League, looking for evidence of non-sport activity. We treat the league as a system of linked elements and we use tools from statistical physics to research some of its properties (Fig. 6). We find which clubs play for other profits by analyzing the statistical situation before and after the matches in which, as ruled by the court, a crime had been committed. This research is dedicated to getting rid of corruption in Polish Football. For example, clubs supported by coal mining companies (Ruch Chorzów and GKS Katowice) are close to each other (Fig. 6).

4. Conclusions

Results of our visualization analysis may develop descriptions of social and economic systems. Methodology of MST diagrams based on correlation coefficients in social and economic phenomena makes data more understandable. It has been already shown in financial markets where it is possible to isolate groups of stocks or currencies that make sense from an economic point of view. Our aim is to offer a mathematical, quantitative perspective, because experts in social science tend to work with unsatisfactory methods (like multi regressions). This is necessary for better understanding of an investigated system. It is also possible to obtain a more reliable comparison between systems (as we presented in this paper). Multiscale dimension mapping methodology of MSTs (scaling time series to 2D graphs) allows us to present relationships in both local and global scale. These methods allow us to focus on new scientific problems. The MST methods have given a better description of people who suffer from thyroid. Thus, it is possible to isolate previously unknown chains of influences in the Minimal Spanning Tree that have direct or indirect impact on the quality of life.

In political studies, we ask a question: Is polarization observed in both regimes (media and real political action) or only in the perception of media? And what does characteristic fingerprints of such a critical event as the Smolensk crash to political ties in general (in meaningful commodity market taxonomy where 'prize' of a political party is an equivalent to opinion polls). While social scientists try to understand mechanisms of political polarization with underlying psychological and political patterns, our hierarchical network analysis with application of natural language processing and text mining, are great supplementary tools for that purpose. The aim of this preliminary exploratory quantitative study was to generate questions and hypotheses, which could be followed by careful qualitative methods in the future.

For the phonographic market the conclusion is that from an economical point of view, pop music does not exist in the way that record companies think (a superstar cluster instead of a genre cluster). Pop music is a term that originally derives from of the word 'popular' and its extension to a music genre is not allowed. We successfully utilized both a qualitative approach (aimed at the so-called conscious consumers) and a quantitative one (with methods of statistical physics, aimed at mass consumers only

Fig. 6. Hierarchical representation of the First Polish League season 2003/2004. Ranking (place in table) and result in game as an average of signal.

weakly related to the phonographic and music industry). Finally, we expect that our research will contribute to the knowledge of dynamic phenomena occurring in complex systems, especially those of society and economics.

References

Buda, A. and A Jarynowski (2010), *Lifetime of correlation*, WN, Wroclaw.

Buda, A. and A. Jarynowski (2013), 'Network structure of phonographic market with characteristic similarities between artists', *Acta Physica Polonica A* **123**, 123.

Grabowski, A. and R.A. Kosinski (2004), 'Epidemic spreading in a hierarchical social network', *Physical review. E* pp. 031908.1–031908.7.

Green, D.G. and E.J. Bossomaier (2000), *Complex systems*, Cambridge University Press.

Jarynowski, A. (2010), *Anomalous interactions in network of Polish Football League*, WN, Wrocław.

Kwapień, J. and S. Drożdż (2012), 'Physical approach to complex systems', *Physics Reports* **515**, 115–226.

Mantegna, R.N. (1999), 'Hierarchical structure in financial markets', *Eur. Phys. J. B* .

Mezard, M., M. Parisi and A. Virasoro (1987), *Spin glass theory and beyond*, World Scientific, Singapore.

Najmrodzka, S. (2012), Badanie jakości życia pacjentów z chorobami tarczycy„ Technical report, Poznań.

Ordanini, A. (2006), 'Selection models in the music industry', *Journal of Cultural Economics* **30**, 183–200.

Oswiecimka, P., J. Kwapien and S. Drożdż (2011), Computational approach to multifractal music, Technical report.

Watts, D.J. and S.H Strogatz (1998), 'Collective dynamics of small worlds networks', *Nature* **393**.